UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

(1)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AFIT/CI/NR 88-143
2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER

DTIC FILE COPY

AD-A197 322

TITLE *(and Subtitle)*: AUDITORY MODELS FOR SPEECH ANALYSIS

5. TYPE OF REPORT & PERIOD COVERED: ~~MS THESIS~~ TECHNICAL REPORT

6. PERFORMING ORG. REPORT NUMBER

AUTHOR(s): MARK T. MAYBURY

8. CONTRACT OR GRANT NUMBER(s)

PERFORMING ORGANIZATION NAME AND ADDRESS: AFIT STUDENT AT: CAMBRIDGE UNIVERSITY UNITED KINGDOM

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

CONTROLLING OFFICE NAME AND ADDRESS

12. REPORT DATE: 1988

13. NUMBER OF PAGES: 30

14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)*: AFIT/NR, Wright-Patterson AFB OH 45433-6583

15. SECURITY CLASS. *(of this report)*: UNCLASSIFIED

15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

DISTRIBUTED UNLIMITED: APPROVED FOR PUBLIC RELEASE

DTIC ELECTE AUG 12 1988 S D

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

SAME AS REPORT

18. SUPPLEMENTARY NOTES

Approved for Public Release: IAW AFR 190-1
LYNN E. WOLAVER Lynn Wolaver 21 July 88
Dean for Research and Professional Development
Air Force Institute of Technology
Wright-Patterson AFB OH 45433-6583

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

ATTACHED

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

ABSTRACT

This paper reviews the psychophysical basis for auditory models and discusses their application to automatic speech recognition. First an overview of the human auditory system is presented, followed by a review of current knowledge gleaned from neurological and psychoacoustic experimentation. Next, a general framework describes established peripheral auditory models which are based on well-understood properties of the peripheral auditory system. This is followed by a discussion of current enhancements to that model to include nonlinearities and synchrony information as well as other higher auditory functions. Finally, the initial performance of auditory models in the task of speech recognition is examined and additional applications are mentioned.

MARK T. MAYBURY

WOLFSON COLLEGE

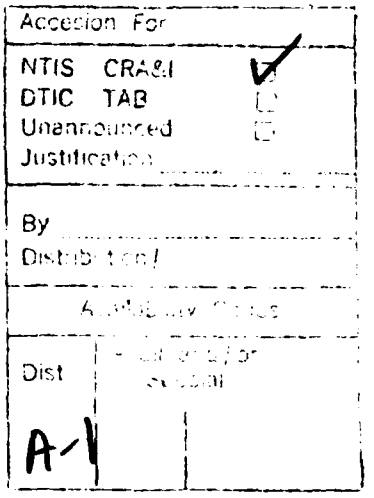
Auditory Models for Speech Analysis

Mark T. Maybury

Cambridge University
Engineering Department

Trumpington Street
Cambridge CB2 1PZ

Wolfson College, 1987



ABSTRACT

This paper reviews the psychophysical basis for auditory models and discusses their application to automatic speech recognition. First an overview of the human auditory system is presented, followed by a review of current knowledge gleaned from neurological and psychoacoustic experimentation. Next, a general framework describes established peripheral auditory models which are based on well-understood properties of the peripheral auditory system. This is followed by a discussion of current enhancements to that model to include nonlinearities and synchrony information as well as other higher auditory functions. Finally, the initial performance of auditory models in the task of speech recognition is examined and additional applications are mentioned.

Contents

1 INTRODUCTION

Current speech recognition systems have demonstrated less than ideal results in speaker independent recognition tasks. Many researchers believe the problem lies in the current spectral representation which does not account for speaker variability and other nonlinearities. Advances in speech recognition attributed to the acoustic theory of speech production suggest that a more robust theory of auditory perception should circumvent the failings of the sound spectrograph and guide the construction of more effective speech understanding systems Furthermore, enhanced auditory models should also support the development of more precise hearing aids as well as aid the design of warning systems, such as alarms. With such applications in mind, it seems only natural to begin by examining the most effective speech recogniser yet developed: the human auditory system.

2 HUMAN AUDITORY SYSTEM

This brief overview of the structure and function of the complex and only partially understood human auditory system is necessarily simplified. The purpose is to review our current knowledge of the ear so that the artificial auditory models for use in speech recognition systems can be evaluated in terms of physiological and functional validity.

2.1 The External and Middle Ear

On a gross anatomical level, the ear consists of three major components: the external ear, the middle ear and the inner ear (see Figure 2.1). The external ear includes the *pinna* and the tube-shaped auditory canal, known as the *meatus*. The pinna, though often ignored, modifies primarily the high frequency of the incoming signal and plays an important role in sound localisation (Moore, 1982). The meatus emphasises frequencies near its 3.4 kHz (Shaw, 1974) fundamental resonance, as well as integral multiples of this frequency [1], and can be modelled by a linear pre-emphasis filter (Klatt, 1982). After travelling down the approximately 2.5 cm meatus, the sound impinges on the *tympanic membrane*, commonly referred to as the eardrum.

Attached to this membrane are three small bones, first the *malleus* (hammer), then the *incus* (anvil) and finally the *stapes* (stirrup), together called the *ossicles*. This *ossicular*

[1] This interestingly is the location of much consonantal and upper vowel formant information (Durrant and Lovrinic, 1984).

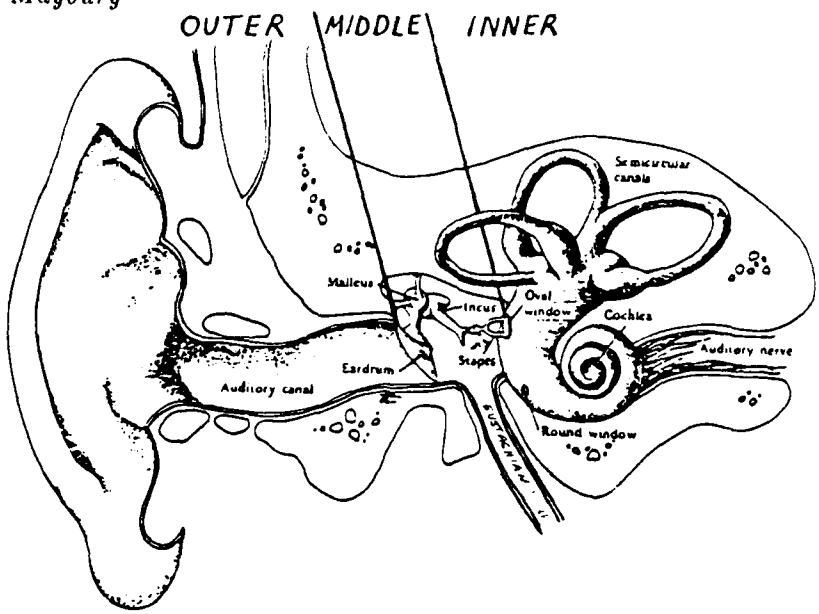
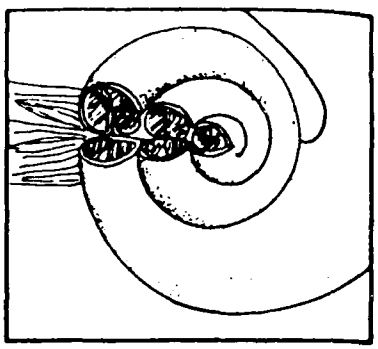


Figure 2.1. Structure of the Peripheral Auditory System, illustrating the outer, middle, and inner ear [Moore, 1982, p. 14].

chain, located in the middle ear, connects the tympanic membrane with the spiral-shaped cochlea and acts primarily as a transformer to amplify the sound (by about 30 dB) so that the incoming wave has sufficient impact on the denser (with respect to air) saline solution in the cochlea (Borden and Harris, 1983, p. 171). The middle ear matches the disparate impedances of air and liquid by two mechanism. First, it compresses the vibration from the large effective area of the tympanic membrane (0.55 cm^2) onto the much smaller oval window (0.03 cm^2) which augments the pressure by 25 dB. Second, the lever action of the ossicles adds several more decibels.

The middle ear performs other functions. First, it attenuates loud sounds of more than 85 or 90 dB by use of the *acoustic reflex* whereby the smallest human muscle, the *stapedius muscle*, contracts to reduce the amplification of the signal. It is thought that the acoustic reflex reduces the audibility of self-generated sounds and causes lower frequency masking of middle and higher frequencies (Moore, 1982). This attenuation may help preserve vowel information normally lost by F1 masking F2 and F3 (Sachs and Young, 1980) . Finally, the eustachian tube maintains air-pressure equilibrium with the atmosphere by providing an air conduit to the nasopharynx.



Human cochlea (sectioned)

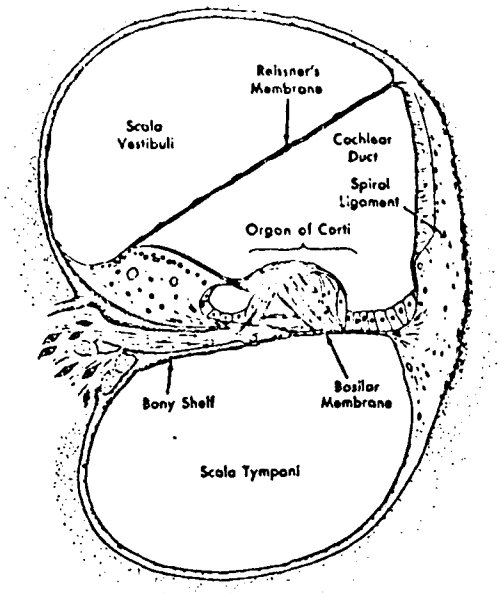


Figure 2.2. Illustration of the Cochlea showing the Organ of the Corti, basilar membrane, and hair cells [Lindsay and Norman, 1977; Borden and Harris, 1983].

2.2 The Inner Ear

Sound travels from the ossicles to the *base* or beginning of the liquid-filled cochlea via the *oval window*. The wave then propagates down a 3.5 cm channel called the *scala vestibuli* through the incompressible cochlear fluids, the *perilymph*, to the inner tip of the spiral, the *apex* (see Figure 2.2). From here it passes through the helicoptrema, returning via the parallel *scala tympani*, eventually being absorbed by the *round window*.

The *cochlear duct* lies between these two channels separated by the two flexible membranes, Reissner's membrane and the frequency-sensitive *basilar membrane*. At the cochlea base, the basilar membrane is narrow and stiff, becoming gradually wider and more flaccid toward the apex. These physiological characteristics make high-frequency response greater at the base and low-frequency response greater in the apex of the basilar membrane (Moore, 1982). Hence the basilar membrane performs a rough Fourier analysis on the incoming signal and maps frequency onto location on the membrane. Von Bekesy (1928, 1942) completed most of the pioneering work concerning vibrations along the basilar membrane. He suggested the travelling wave theory as well the basilar membrane stiffness qualities.

However, the true auditory sensor is not the basilar membrane but the attached *Organ of Corti* which contains the approximately 23,500 inner and outer hair cells separated by an arch known as the tunnel of Corti (Seneff, 1985). The undulation in perilymph is thought to cause a shearing motion between the basilar membrane and the tectorial membrane

(which lies above the hairs) which in turn bends the hairs [2] and causes electrochemical excitation on a given auditory nerve fiber. These fibers carry the signal along the eighth cranial nerve to the cochlear nucleus, the beginning of the central auditory system.

2.3 Higher Auditory System

Physiologically, the pathway to the auditory cortex from the cochlea passes through the cochlear nucleus, the superior olive, the inferior colliculus, and finally the medial geniculate. As the hierarchy in figure 2.3 illustrates, these parts are interconnected in a non-trivial manner. Recent experimentation offers some insight into the role each of these areas plays in speech perception.

"Octopus cells" found in the posteroventral cochlear nucleus have been stimulated by tone bursts at CF and show a sharp response at first, then little or no activity. One suggestion is that both excitation and delayed inhibition is at work. When stimulated by a click train these cells appear to encode stimulus frequency by firing in period with the clicks, up to a critical value. (Godfrey *et al.*, 1975). Some binaural processing appears to occur in the superior olivary nucleus as well as in the inferior collicus and perhaps in higher levels also (Lindsay and Norman, 1977). Møller (1982) has shown a "neurophysiological basis for the perception of complex sounds" in the inferior collicus.

In the auditory cortex, Whitfield (1967) noticed that only sixty percent of the neurons respond to pure tones – but in a complex manner. When a tone is present, some neurons are excitatory; others are inhibitory. Others respond only when a tone is on or off and still others when the tone is turned on and off. Some neural units act as frequency sweep detectors (Wattfield and Evans, 1965, discussed in Moore, 1982), responding only to frequency changes; others act as interactive cells, inhibiting response to tones at other frequencies when a tone at a specific frequency is present. Unfortunately, neural pathways between the various nuclei are even less well understood (Moore, 1982).

Examining cortical neurons, Abeles and Goldstein (1972) found three types of tuning properties: narrow, broad and multirange. Hrugge and Merzenich (1973) found that

[2] Recent micromanipulation on the Organ of the Corti in the guinea pig (Flock, 1982, p. 47) concluded that "sensory hairs are tuned and are likely to contribute to frequency selectivity of the Organ of the Corti through their mechanical coupling to the tectorial membrane." Furthermore, the sensory hair stiffness was found to be twice as large in the excitatory direction (moving toward the centriole) than in the inhibitory way, which could account for some well known non-linear cochlear effects.

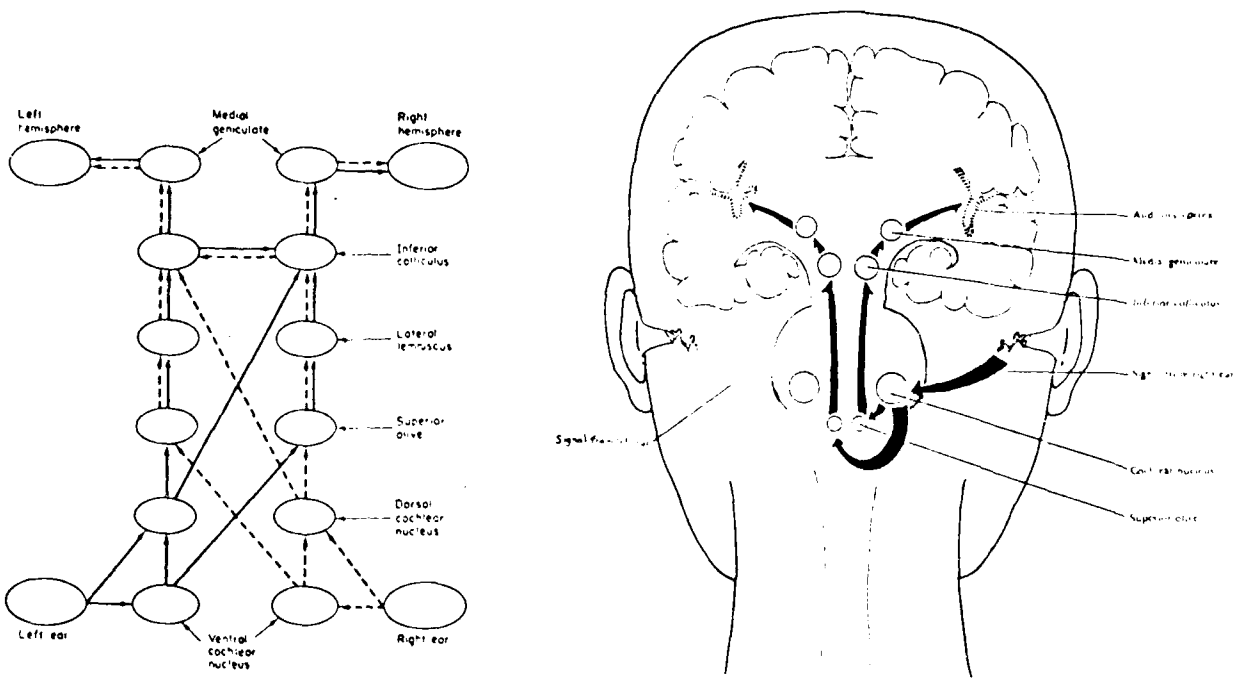


Figure 2.3. Illustration of the Higher Auditory System
[Moore, 1982, p. 36; Lindsay and Norman, 1977, p. 241].

unanaesthetised monkeys had neural units sensitive to intensity and interaural tone differences. Wollberg and Newman (1972) examined cortical neurons in squirrel monkeys and found cells which responded to many different sounds but others which responded to only one call. These experiments suggest that neural cells are sensitive to acoustic feature in a sense similar to feature detecting cells found in the retina. Without additional evidence, however, it would be irresponsible to claim a one-to-one correspondence between animal experimental results and human audition. More importantly, for our purposes, it is nei-

ther feasible nor clear how to incorporate our current knowledge into a computer model for automatic speech recognition.

3 PLACE AND SYNCHRONY

As previously noted, the peripheral auditory system converts the incoming signal into a pattern of action potentials along the auditory nerve. Currently two perhaps complimentary proposals account for this transfer at the cochlea: the *place theory* and the *temporal theory*.

The place theory (Helmholz, 1867) states that the tuning characteristics of individual nerve fibers vary according to their distance from the apex of the cochlea[3]. Fibers close to the base tune in to higher center frequencies (CF) and those nearer the apex respond better to lower CF. The final result is a "spike train of action potentials with a non-homogeneous Poisson-like distribution." (Seneff, 1985) Even without an external signal, nerve fibers exhibit a spontaneous or background firing of about 150 impulses/sec. Firing appears to occur independently of adjacent acoustic fiber activity, hence offering improved sampling accuracy (Johnson, 1980 in Seneff, 1985). Fibers also adapt to stimuli by decreasing their response rate after the onset of a signal (Smith and Zwislocki, 1975) This decrease in sensitivity is sharpest immediately following stimulus onset, followed by a slow decay.

Temporal excitation patterns (Zwicker and Feldtkeller, 1967), on the other hand, characterise the way in which nerve fibers synchronise their firing to periodic wave patterns. This feature is particularly evident at low frequencies (below 5 kHz) where the phase-locked neuron fires in synchrony with the cycle of the local dominant frequency (Lindsay and Norman, 1977). Fibers with high center frequencies (CF), on the other hand, tend to phase-lock with the envelope of a stimulus. It has been suggested (Delgutte, 1980; Delgutte and Kiang, 1984) that this "envelope synchrony" could aid pitch detection since the envelope responds to the fundamental period of a stimulus.

Ernest Glen Wever (Borden and Harris, 1983, p. 175) similarly noted that at low frequencies nerve fiber displacement was not sharp, but the number of cycles/sec in the stimulus corresponded to a similar number of nerve fiber impulses/sec. At high frequencies

[3] This theory suggests the cochlea is best represented as a bank of overlapping near bandpass filters. Schroeder, Atal, and Hall (1979) suggested that each band pass filter corresponds to a 1.5mm section of the basilar membrane which enervates approximately 1200 neurons.

the number of impulses/sec was too high for a single fiber to accurately record periodicity information. He suggested a "volley theory" whereby neurons cooperate at high frequencies in order to capture the rapidly appearing wave cycles. He also noted a relationship between intensity and relative rate of spikes, limited by a threshold (see section 4.5).

A "period histogram" is a method of representing these response patterns of fibers to a periodic input signal, such as a sine wave (see Figure 4.1). A plot can be obtained by measuring nerve impulse responses occurring within segmented periods of the stimulus (Seneff, 1985). This detailed pattern of neural activity with respect to time could prove useful in frequency detection.

4 NEUROPHYSIOLOGICAL AND PERCEPTUAL EXPERIMENTS

A number of physiological and perceptual experiments have been undertaken in an attempt to capture the essence of the auditory system. Experimental results have led to the assumption that the periphery consists of a filter bank followed by a "nonlinear time-dependent amplitude compression, including half-wave rectification" (Seneff, 1985).

4.1 Peripheral Auditory Response

The belief that the auditory periphery acts essentially as linear filter (complicated at higher levels by nonlinearities such as half-wave rectification, saturation and adaptation) arose from results of three methods which achieved similar but unequal results (Seneff, 1985). The most natural as well as inaccurate method involves determining filter shape from perceptual masking tests performed on humans. In addition, vibrations of the basilar membrane were measured using the accurate Mossbauer technique (Johnston and Boyle, 1967) and laser interferometry (Khanna and Leonard, 1982). Finally, "neurophysiological tuning curves" (NTC) were derived from direct measurements of nerve fiber responses at various locations on the basilar membrane. The results of these methods are used to construct what Helmholtz in 1863 described as a bank of overlapping, linear bandpass filters – the implementation of the place theory.

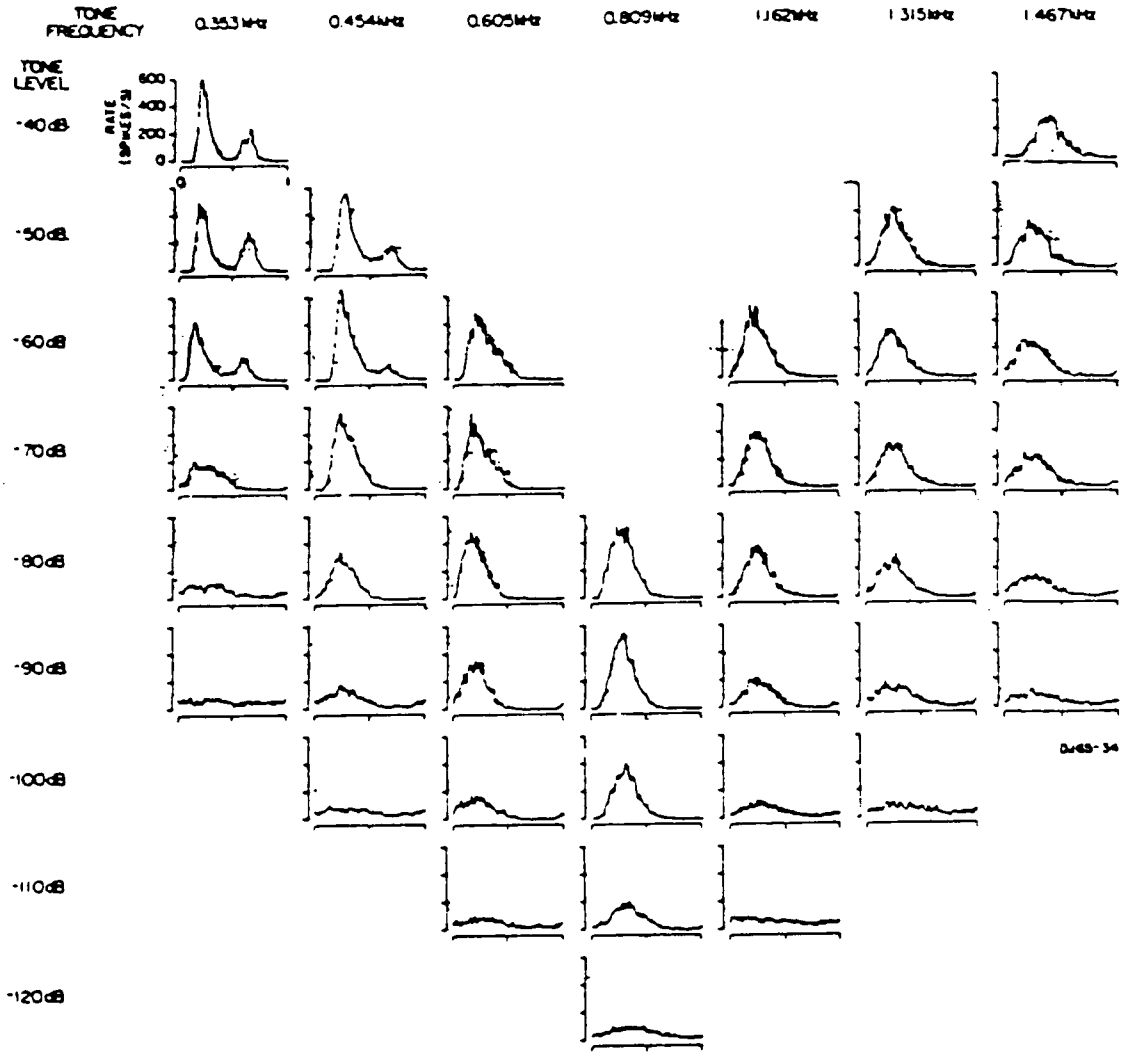


Figure 4.1. A period histogram illustrating the nerve firing patterns over time, represented in spikes/sec. Each histogram displays 15 sec responses characteristics to single tones. Sudden activity following tone onset is followed by a rapid decay to a steady state. [Johnson, 1980].

4.2 Psychophysical Tuning Curves (PTC)

In 1940, Fletcher formalised the frequency resolution of the basilar membrane with his "critical band" concept. By masking pure tones with bandpass noise, he observed that increases beyond a "critical bandwidth" made little difference in the amplitude level just necessary to mask the tone. He then characterised the peripheral auditory system

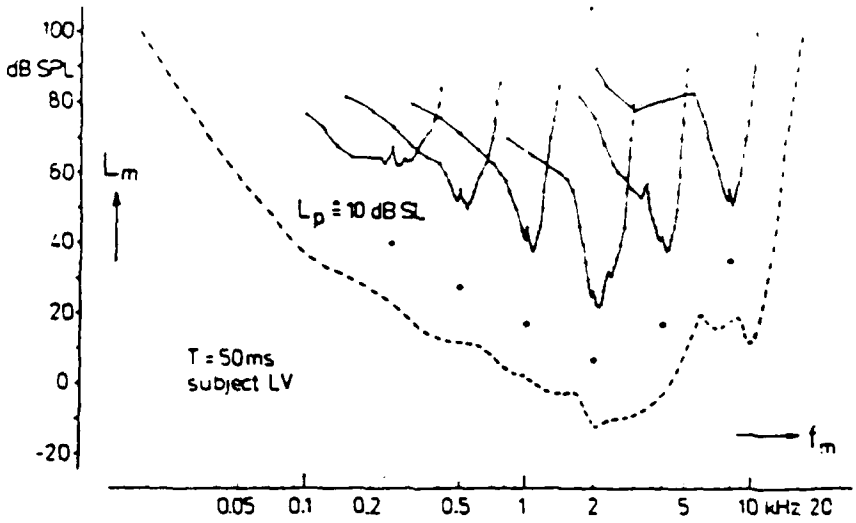


Figure 4.2. Psychophysical Tuning Curve (PTC) obtained from masking experiments [Vogten, 1974].

as a bank of bandpass filters with continuously overlapping frequencies of critical bands, varying with frequency.

Houtgast (1977) and Pick (1980) used an asymmetric rippled noise masker and similarly found that the auditory filter is nearly symmetric at moderate noise levels (Patterson and Moore, 1986). Patterson and Moore (1986) offered a notched formula and demonstrated that the filter varies with age. Also, since the filter has asymmetry at high and low levels (Paterson and Nimmo-Smith, 1980), they put a notch in the noise both symmetrically and asymmetrically about the signal (Patterson and Moore, 1986).

These masking experiments avoid sinosoids since a sinusoidal signal combined with a sinusoidal masker can cause "beats" due to regular fluctuations in the combined envelope (Patterson and Moore, 1982). Nevertheless, masking is difficult because it is subject to both swamping or suppression phenomena (Moore 1982). In general, psychoacoustic masking experiments are the least reliable method of obtaining tuning curves.

4.3 Neurophysiological Tuning Curves (NTC)

A more direct way to determine auditory filter shape is to measure neural response to stimuli. Generally nerve fibers are stimulated by pure tones and measured for responses above spontaneous levels (Kiang *et al.*, 1965 from Seneff, 1985). These results must be carefully interpreted since neurophysiological experiments performed on cats together with psychophysical experiments performed on humans have shown that nerve fiber responses are physiologically vulnerable (Moore, 1982).

Von Bekesy (1960) pioneered yet another method of obtaining filter shapes by directly measuring the motion of the basilar membrane using optical techniques. Johnstone and Boyle (1967) obtained finer tuning curves by applying one of these methods, the Mössbauer technique[4], to obtain a "transfer function of the cochlea" which plotted basilar membrane displacement versus stapes displacement.

When compared with the 100db/octave slope of the neurophysiological tuning curves (Evans and Wilson 1973), the slopes of Johnstone and Boyle's curves are much shallower at 13 dB/octave at low frequencies, and slightly steeper at 105dB/octave at high frequencies. One suggestion to account for these dissimilarities is that not the amplitude but the velocity or acceleration of the basilar membrane characterises the response of the hair cells (Hall, 1980). Another suggestion is that a "second" physiological filter acts between the basilar membrane and the neurons.

Yet another explanation is that lateral inhibition of neurons (strong inputs suppress weaker neighbours) results in excitation patterns boardered by inhibited areas (von Bekesy, 1960). Zwislocki (1965) calculated what von Bekesy called neural inhibitory units to correspond to the critical bandwidth. Robertson and Manley's (1974) observations of the effects of hypoxia on the sharpness of the tuning curves lend credibility to this explanation (Moore, 1982).

---

[4] This technique first places Gamma rays on the membrane with an absorber nearby. The movement of the basilar membrane can then be measured to an accuracy of 1mm/sec (Greene 1976), relying on the Doppler effect.

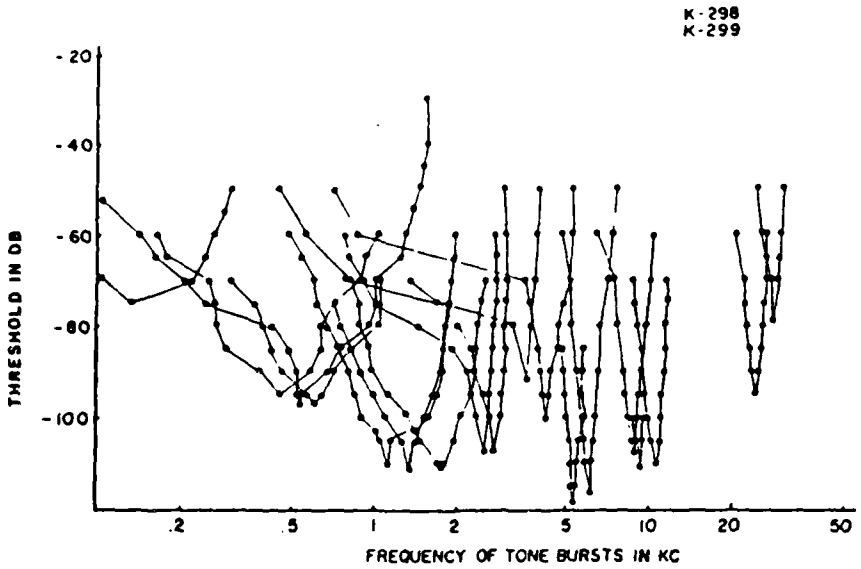


Figure 4.3. Neurophysiological Tuning Curve (NTC) obtained from direct measurement of nerve responses of cat fibers [Kiang *et al.*, 1965].

4.4 PTC's and NTC's

Recently, more sophisticated experiments have been devised to determine the filter shape, including varying the center frequency of the noise masker. Assuming that the masker level necessary to disguise the signal was proportional to the response level of the nerve fibers, Vogten (1974) obtained a PTC's which were similar in shape to the NTC's (Seneff, 1985). (Compare Figures 4.2 and 4.3.)

One must view this conclusion with caution, however, as PTC's are determined with a single masker and signal presented simultaneously while NTC's result from stimulating a neuron with one tone at a time (Moore, 1982). Furthermore, humans are known to perform "off-frequency listening", attending to the filter which gives the best signal-to-noise ratio (SNR) rather than that which is centered on the signal. The effect is a sharper tip in the PTC than if just one filter were involved (Johnson-Davies and Patterson, 1979; O'Loughlin and Moore, 1981). Patterson (1976) prevented off-frequency listening by measuring the threshold of the signal as a function of the width of a notched noise masker (Moore, 1982). The level of noise necessary to mask the signal is measured, realising that asymmetries between high and low bands will not appear since the noise notch is symmetrical. To date, however, this is an area of further research.

4.5 Intensity

The linear filter must include phase and gain information in addition to tuning curve shape in order to accurately characterise auditory response. Von Bekesy (1960) and more recently Rhode (1971) measured phase response in the basilar membrane and in the final analysis found a linear relationship below CF which became increasingly steep approaching CF (Seneff, 1985 figure 2.4). Allen (1983) found similar results from measuring neuron firings. Wever's (1970) previously mentioned "volley theory" [5] could also account for intensity coding.

4.6 Auditory Nerve Responses to Stimuli

Generating NTC's from experiments with single tones fails to capture the inhibitory effects of multiple tone stimulus. For example, masking can occur both simultaneously as well as sequentially (ie forward and backward masking). More sophisticated experiments should, then, yield a more accurate NTC.

Johnson (1980) characterised nerve fiber response as a function of the frequency and amplitude of the stimulus. By computing period histograms he demonstrated a peak-splitting phenomena with low frequency, high intensity isolated tones. He then suggested this observation to be a result of nonlinearities in the cochlea which introduce a prominent second harmonic.

By examining short-term adaptation of guinea pig nerve fibers to "tone pedestals", Smith and Zwislocki (1975) found that adaptation is essentially linear. However the rate of decay was more rapid immediately following tone onset.

Sachs and Kiang (1968) demonstrated nonlinearities in the system by measuring nerve fiber responses to two-tone complexes. They found "two tone suppression" (where a low frequency tone masks a higher frequency tone, or vice-versa) to manifest itself in neuron fiber rate responses. The response can even decrease with increased energy, if a second tone (near the edge of the fiber tuning curve) is added to a tone at the characteristic frequency (Seneff, 1985).

Rose (1974) derived a formula to describe two-tone stimulus response which matches response histograms "with half-rectified sums of two sine waves at the tonal frequencies,

[5] Recall that each cycle of the signal elicits firing in different nerves so that at high frequencies the combined pattern represents all cycles in the stimulus.

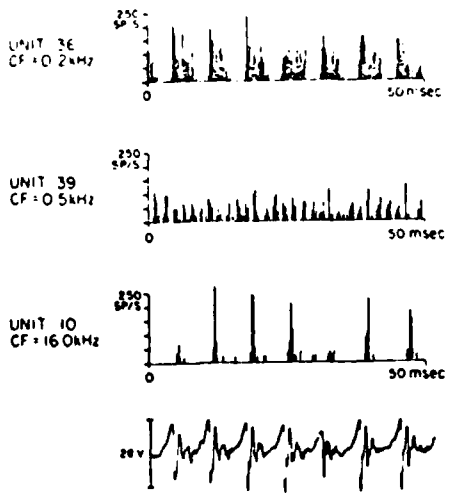


Figure 4.4. Cat Auditory Neuron Response to a segment of vowel /ae/.
[Borden and Harris, 1983, p. 175]

with appropriate amplitude and phase" (Seneff, 1985). Hall (1980) recently suggested that hair cells may be sensitive to velocity or acceleration of the basilar membrane and two tone suppression could be a suppression effect by both the already tuned motion of the basilar membrane as well as the "second filter" (Seneff, 1985).

In other experiments, forward masking (presenting a mask signal followed by silence then the stimulus) was addressed by Harris and Dallos (1979). No surprisingly, they found an exponential relationship between the response rate and time delay between masker and stimulus. They concluded that reduced nerve fiber response is a definition of short-term adaptation.

4.7 Rate and Phase Locking

Some interesting experiments have shed light on the rate and synchrony theories of neural firing. These began with experiments in neural response to speech stimuli. Sachs and Young (1979, 1980) measured the response of a cat's auditory neurons to steady state synthetic /ϵ/, Miller and Sachs (1981) measured response to synthetic CV (eg /ba/ /da/) and Delgutte (1980) examined transition and steady-state responses for fricative-like and vowel-like stimuli. The results were consistent with previous experiments (eg adaptation, two-tone masking). Figure 4.4 illustrates a typical result: the neurons in the auditory nerve of a cat maintain a fixed temporal relationship with the presented vowel /ae/.

Sachs and Young (1982) pursued this further by examing both rate response and histograms of excitation patterns for the purpose of identifying synthetic vowels. They found that for high amplitudes, the rate response characteristics lost formant information as a result of saturation and two-tone masking effects. This gives credibility to the theory that patterns in the excitation, specifically synchronicity information, can contribute to a clearer formant picture.

Sachs and Young then developed the "Average Localised Synchronised Rate" (ALSR) which measured nerve fiber energy in response to specific frequencies. Applying this to format motion detection in synthetic /ba/ and /da/ (Miller and Sachs, 1983) they achieved better results than detection based on firing rate alone.

Delgutte (1980) stimulated cat auditory nerve fibers with CV tone and noise bursts as well as steady state vowels and found the rate jumped sharply at onset, decayed rapidly for 15-20 ms, then slowly decayed to a steady state. He also examined "post adaptation" effects, where a stimulus is preceded by a long adaptation signal. As might be expected, increasing the adaptation signal corresponds to a decreased response rate in the subsequent stimulus. In other experiments he found responses at vowel onset in /ba/ compared to /ma/ exhibited sharper peaks onsets presumably because the /m/ adapted to low frequency fibers more than the /b/.

When he examined responses to single-formant vowel-like stimuli, Delgutte found considerable periodicity represented in the nerve firing patterns, at low frequencies. Probably due to saturation effects, however, the periodicity of the envelope of the signal was not present in the fibers tuned to the formant frequency at higher amplitudes.

In sum, both synchronicity and rate information is necessary to resolve frequency. This seems to make functional sense as redundant mechanisms cooperate to yield greater accuracy.

5 SPEECH PROCESSING BASED ON AUDITORY MODELS

Several models of the human peripheral auditory system have been suggested (Zwicker *et al.* 1979, Searle *et al.* 1979, Dolmazon *et al.* 1977; Dolmazon 1982, Chistovich *et al.*, 1974; Chistovich *et al.*, 1982) for application to speech processing. Most include the following functions which we have seen in our overview of the peripheral auditory system (Klatt, 1982).

1. A linear preemphasis filter to model the mid-frequency boost provided by the external ear canal and middle ear.
2. A set of more-or-less linear bandpass "critical-band" filters spaced equally along a Mel or Bark frequency scale to model basilar membrane mechanics.
3. Half-wave rectifiers to model the transformation that takes place at the hair cells.
4. Low-pass filters with rather short time constants, at least in the high-frequency channels.
5. Lateral suppression circuitry to sharpen peaks seen in the output spectra (Houtgast, 1977; Houtgast and van Veen, 1982; Sachs and Kiang, 1968).
6. Partial adaptation of filter outputs via time derivatives or other computations that emphasise onsets and perhaps offsets (Delgutte, 1982).
7. A log transformation of filter outputs to approximate the phone scale of loudness.

Certain other functional properties have been added to model nonlinearities discussed earlier. For example the low frequency attenuation of the middle ear at high intensities, the nonlinearities in the basilar membrane with respect to complex tones and masking (Allen, 1980) as well as excitation decay patterns as a function of signal duration (Zwicker *et al.* 1979) are refinements of the above core model.

Unfortunately, the audio spectrogram produced as a result of implementing the above set of linear properties is less than satisfactory (Chistovich, 1982) (the 100 Hz bandwidth of the critical band filters are too narrow). Physiological and psychophysical results argue against a larger bandwidth. Furthermore, a larger low-frequency bandwidth may actually reduce intelligibility (Klatt, 1982).

5.1 Toward an Auditory Spectrogram

The broad band sound spectrograph used in speech research for the past forty years could mislead us by providing information about incorrect perceptual cues. For example, voices of women and children have higher fundamental frequencies, are more breathy, and, largely due to shorter vocal tracts, have higher formants. Yet humans find little perceptual difficulty in normalising for sex and age. This suggests we need a better representation than the current frequency-amplitude-time pattern.

Auditorily transformed spectra could be superior input to speaker- independent speech recognition systems (Bladon, 1985). Carlson and Granström (1982) compared alternative methods for spectograph generation and found a Bark/Phon [6] scale to most accurately relate to psychoacoustics and speech perception.

Syrdal and Gopal (1986) built on the theory of Potter and Steinberg (1941) that vowel recognition is based on spatial patterns of excitation in the peripheral auditory system regardless of location on the basilar membrane. They were able to show that vowel classification after linear indiscrimination analyses were significantly more accurate using the bark scale than the Hertz scale. In addition, the bark scale gives added performance by providing inherent speaker independent normalisation while preserving linguistic validity. Furthermore, bark difference can be used to link acoustic and phonetic features.

5.2 Synchrony Models

While our peripheral auditory model may provide better input than traditional spectrograms, it is unclear what occurs at higher levels. It may be that central auditory mechanisms perform spectral smoothing but currently there is conflicting evidence on this proposal (Klatt, 1982). Unfortunately, the constraints of the peripheral auditory system will not provide a representation which will normalise interspeaker variability (Bladon, 1985).

However we can exploit additional information concerning interspike intervals, at least below 5 kHz. This may be useful in pitch detection (Sachs, 1982). Sachs and Young (1980)

[6] Zwicker's (1961) bark scale, named after the inventor of the unit of loudness, Barkhausen, characterises the critical bands on a non-linear scale based on the logarithm of the frequency. One bark is equivalent to the width of a critical band on the frequency scale.

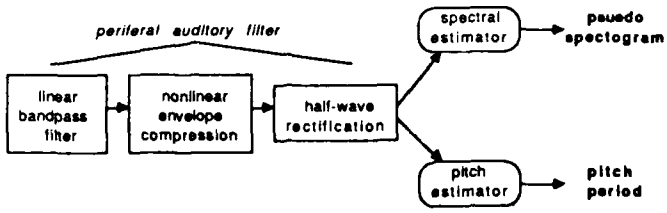


Figure 5.1. Overview of Seneff's Auditory Model

examined patterns of interspike intervals for nerve fiber responses to vowels and found saturation at 60 or 70 dB SPL, but invariant patterns (Klatt, 1982).

Carlson and Granström (1982) have suggested a model (DOMIN) that instead of measuring synchrony to center frequency of the nerve fiber (eg Srulovicz and Goldstein), finds which frequency dominates each point along the basilar membrane. These can be superimposed on the traditional spectrogram (Carlson and Granström, 1982). The authors used these plots as templates and found if used in a traditional template matching recognition scheme they obtained better accuracy with DOMIN templates than with smoothed DFT representations. In sum, a dominant frequency estimation is appealing because of the demonstrated perceptual importance of formant frequency locations in recognition of vowels (Klatt, 1982).

Seneff's synchrony model (1985, 1986), based on simple and psychologically motivated operations, offers yet another alternative to the traditional temporal and place-based models of pitch perception (see Figure 5.1)[7]. She uses Generalised Synchrony Detectors (GSD) which take input from a peripheral auditory model of 32 filters with critical bandwidths (with a range from about 200 to 2700 Hz) and output pitch and formant structure. The algorithm begins with two signals, the original (u) and the same signal delayed by τ (v). An envelope of the two is constructed from the low bandpass of a sum and difference of u and v (ie $\sum u + v$ and $\sum u - v$). A small value, δ, representing the spontaneous rate of neural discharge, is subtracted from these envelopes. The sum envelope is divided by

[7] Recently she has extended her model to include short term adaptation as well as a linear low-pass filter to simulate loss of synchrony at high frequency nerve firings.

the difference envelope which is then passed through a soft limiter (eg arctan function) (Seneff, 1985).

A "pseudo spectrum" can be obtained by plotting the GSD's output as a function of the center frequency. Generating this has the desired property of sharpening peaks at formant frequencies (Seneff, 1985). A pitch estimator sums the peripheral level outputs from the thirty-two critical band filters to obtain a single "pitch waveform". Then a series of GSD's at delays covering the range of the human fundamental frequency are applied to this pitch waveform. Finally, a plot of the output as a function of the delay period yields a "pseudo autocorrelation" whose first prominent peak corresponds to the fundamental frequency. This information of pitch and format structure could be useful for speech recognition if carefully incorporated into a more complete auditory model. In particular, a combination of both envelope information to distinguish between broad acoustic classes together with synchrony information to obtain fine distinctions within on category, could yield better speech recognition performance.

While Seneff reported no speech recognition results, Hunt and Lefebvre (1986) applied a variant of the GSD model to digit recognition using dynamic time warping. After altering the GSD filters to reflect psychophysical test results, they found that a model combining the GSD output together with the input to the GSD achieved better performance than the individual inputs, particularly for noisy signals. The GSD algorithm alone, for example, is insensitive to voiceless sounds, voicing onsets, and rapid formant transitions. The combined approach was tested in a connected-word recognition model (with a one-pass dynamic programming algorithm) in a fighter plane and a helicopter. The system suppressed periodic and aperiodic noise, which seems to suggest that a combined approach yields better performance.

5.3 Central Processing Models

Several central auditory models have been proposed which process the spike train of action potentials to perform specific tasks such as sound localisation (Lyon, 1983). Jeffress (1956) proposed a binaural localisation model where the "mechanism receives impulses from corresponding filter sections of the two ears and delays them progressively by small increments, either by means of fine nerve tissue with a slow conduction rate or by a series of synapses." Colborn (1973) detailed this theory and Lyon (1983) implemented a computational model which cross-correlated similar fiber frequency responses from opposite ears. Lindemann (1983) extended this to incorporate lateral inhibition which sharpened the correlation peak. He also demonstrated some similarities with psychoacoustic phenomena (Seneff, 1985).

Sachs and Young (1982) also did further processing on spike train potentials but with the goal of enhancing formant information. Their methodology consisted first of processing speech through a large population of cat nerve fibers to generate response histograms. Then they processed the histograms to enhance spectral peaks. They found that synchrony information in addition to firing rates was necessary to sharpen merging formant peaks. Their technique computed an ALSR (Average Localised Synchronised Rate) at pitch harmonics in the frequency range.

Richard Lyon (1982, 1983, 1984, 1985, 1986) has developed a highly sophisticated model of the auditory system that is being used in automatic speech recognition (ASR) systems. His computational model serves as a good basis for new auditory theories. After initially concentrating on a computational model of binaural localisation and separation, he extended his model to incorporate knowledge from hearing research. He has reported a system where a filtering model of basilar membrane motion drives a hair cell model, the output of which drives a primary auditory neuron model. The neurons are represented by a "leaky-integrate-to-threshold model with a refractory period" (Lyon, 1984, p. 36.1.1). In the model, each hair cell, or filter channel, incorporates several neurons.

He found that the time-of-firing information obtained from this computational model (unlike the less reliable firing rate versus place representation) contained most of the important speech cues: formants, pitch, direction, etc. Using this cochlear and neural representa-

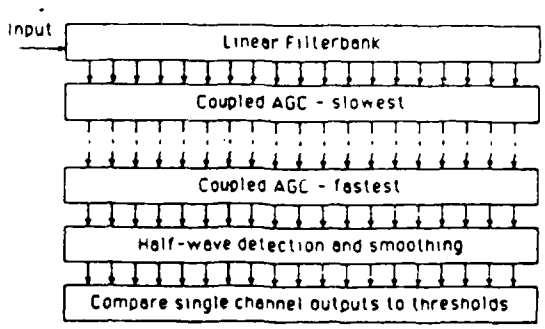


Figure 5.2. Lyon's Cochlear Model,
[Lyon and Dyer, 1986, p. 1975]

tion, he implemented additional models of pitch perception, binaural direction perception and sound separation.

Related work on VLSI design in support of Lyon's computational model has yielded a *multi-serial signal processor (MSSP)* (Lyon and Lauritzen, 1985). This custom bit-serial multiprocessor is intended for real-time implementation of cochlear filtering, compression, hair cell models and several primary auditory neuron models per channel.

In some interesting experimentation with his cochlear system (see figure 5.2), Lyon explains some results of psychoacoustic data. For example, he was able to account for the fact that neural timing curves are sharper than the mechanical transfer function of the basilar membrane as a side effect of a multi-channel automatic gain control (MAGC). This is particularly interesting since no "second filter" is necessary for sharpening. Furthermore, he obtains a "cochleagram" by adding a stage that uses phase to emphasise resonances and spectral peaks. In experiments with 112 adult speakers uttering the numbers one through nine, he found that spectral tilt and loudness variations are de-emphasised by MAGC. In addition, MAGC de-emphasises formant amplitude and bandwidth while retaining clear formant location. These results are important characteristics for performing independent speaker recognition.

5.4 In Search of a Spectral Distance Metric

As we have seen, current auditory models pass the original speech signal through an overlapping sequence of critical band filters, then further process it to obtain a perceptually meaningful representation. This representation generally contains information such as timbre, pitch, loudness and roughness, although researchers fiercely debate both the methods for calculating these parameters as well as their descriptive completeness (Klatt, 1986). Many transformations have been suggested (Klatt, 1982). Unfortunately, our knowledge of the auditory periphery mechanism offers little information on how we are to perform phonetic processing of speech.

Current experiments are aimed at determining the relationship between spectral changes and perceived phonetic changes. Klatt (1982) used pairs of synthetic vowel-like and fricative-like sounds to measure perceived phonetic changes by varying formant frequency, formant bandwidths, spectral tilt and filtering passband/stopband. Similar experiments (Bladon and Lindblom, 1981; Carlson, Granström and Klatt, 1979) found that formant frequency changes are by far the most important perceptual cues.

Results from these experiments should guide selection of a spectral distance metric. This metric – to be used in a traditional template matching scheme – must possess three essential properties: uniform scoring, monotonic response to increasing phonetic distance, and sensitivity to phonetically relevant acoustic cues.

Klatt (1982) suggested a "spectral slope metric" which compares the slope difference between the unknown and the referent along the frequency axis. He found that a weighting function was necessary but this proved to be oversensitive to spectral smoothing of input.

An alternative metric, the "dominant frequency metric" (Carlson and Granström, 1982) starts with a histogram of dominant frequencies instead of a spectra. Unfortunately, this is sensitive to spectral tilt and formant amplitude variation. In short, current distance metrics are unsatisfactory and require more perceptual experimentation to characterise the relevant aspects of phonetic distances.

5.5 Auditory Model Performance

Pattern matching techniques involving referent and unknown discrimination using a distance metric usually involve linear prediction spectra as input (Itakura, 1975). But the lpc weaknesses in characterising fricatives and nasals suggest a new representation would yield better results. The Mel frequency scale and critical band filter perceptually characterise the input signal much more effectively (Klatt, 1982). It is unclear, however, what distance metric would be used. This is critical for effective recognition.

An early performance test of auditory models (Alinat, 1979) involved phoneme recognition by a cochlear model using a bank of 96 bandpass filters with CF based on psychoacoustic data. The recogniser classified vowels, fricatives and stop consonants using features as the base level recognition of phonemes. The test data were two hundred isolated words, uttered clearly and slowly from ten different speakers who spoke twenty words each. Recognition for consonant was 96% and for unvoiced fricative consonants was 100%. Voiced fricative detection was less impressive: only 65% were detected as fricatives and 35% were detected as m, n, l, or voiced stops. 18% of the classification errors occurred in nasal vowel formants. Nevertheless, the consonant and unvoiced fricative results were encouraging.

More recently, Blomberg *et al.* (1986) tested auditory models as front ends to a speech recognition system. They obtained the following accuracy for an isolated word-recognition technique using pattern matching and dynamic programming. The metric was the Euclidean distance.

MODEL	VOWELS		CONSONANTS		MEAN
	Male	Female	Male	Female	
FFT	99	96	95	97	97
BARK	99	95	92	95	95
PHON	96	91	88	91	92
SONE	91	93	83	83	87
DOMIN	99	99	90	90	94

While FFT offers the best overall performance, it is interesting to note the relative success of the alternatives. The DOMIN performs best in vowel recognition. The bark

alternative appears the most promising. Unfortunately, these results are vulnerable to lack of well-understood higher order processing.

In fact, Blomberg *et al.* (1982) performed a similar isolated word-recognition test and achieved the differing results:

MODEL	VOWELS (45 stimuli)		CONSONANTS (90 Stimuli)	
	Errors	% Errors	Errors	% Errors
FFT	6	13	9	10
BARK	4	9	6	7
MASK	6	13	20	22
PHON	9	20	20	22
SONE	10	22	33	37
DOMIN	2	4	30	33
DOMPHON	3	7	15	17

While it is difficult to make clear conclusions from this data, it seems evident that here the bark representation offers the best performance. It appears that as model complexity increases performance drops: MASK adds a psychoacoustic filter to the BARK model, PHON adds a loudness measure to this while SONE adds a perceived loudness measure, and DOMPHON combines both DOMIN and PHON models. This is not the case with the DOMIN model which performs superbly on vowels but fails miserably with consonant detection. This is not surprising if one recalls the spectral peak emphasis (significant in vowel discrimination) inherent in DOMIN. Unfortunately, the DOMIN, like the PHON and SONE models, separates low frequency harmonics which hinders performance. Furthermore, the DOMIN model fails to capture loudness information which is key to recognition, particularly if the speech has the same place of articulation. Nevertheless, DOMPHON showed interesting performance which suggests that combining different forms of analysis aids in the recognition process (Blomberg *et al.*, 1982).

Unfortunately, these experiments fail to support auditory models as a better alternative to traditional speech recognition approaches. At the same time, one cannot disclaim auditory models on these results alone as these models all certainly fail to capture the nuances of the true human auditory system (eg Seneff, 1986; Lyon, 1984). Results of combined modelling approaches seem to indicate a need for many interactive processes.

Also, future modelling of the parallel processing capacity of the higher auditory system should yield more significant results in speaker independent recognition tasks. Furthermore, top-down syntactic, semantic and pragmatic constraints as well as world knowledge could guide recognition. In conclusion, we must wait for further developments in auditory modelling to determine its fate in speech recognition.

6 OTHER APPLICATIONS

In addition to promising better speech processing, auditory models – and the increased knowledge of the human auditory system which they imply – offer much to hearing aid fabrication and alarm system design. For example, equiloudness contours help create more effective (and safer) alarm systems. In hearing disfunction, simulation of impairment can aid in restoring gain, frequency response, and equiloudness contours for often highly idiosyncratic auditory disorders.

Leijon (1982) has developed a semi-automatic "fitting system" which collects relevant audiological information to facilitate digital hearing aid programming. The technique relies on articulation theory, loudness density spectrums, and physical models. These theories aid in the correction of cochlear impairments such as loss in frequency selectivity.

7 CONCLUSION

In closing, we recall that current auditory models for speech processing incorporate a model of the peripheral auditory system together with invented components which exploit interspike interval patterns. These are motivated by the phase and synchrony theories of basilar membrane frequency analysis. However, physiological results cannot yet dictate choice of central processing models.

With regard to speech recognition, we can evaluate the success of auditory models by observing signal-processing results (eg format peak enhancements), robustness to speaker and phonetic context variation, and prediction of psychophysical performances. Preliminary tests in the traditional pattern matching paradigm for speech recognition fail to demonstrate the superiority of auditory models over the standard FFT. This could be due to insufficient central auditory system modeling and inadequate distance metrics. Nevertheless, the auditory models have shown promise in solving the invariance problem and

account for some psychophysical phenomena. Furthermore, they have useful applications, such as in hearing aid design.

In conclusion, auditory models have not yet demonstrated the performance levels necessary to replace current technology used in speech recognition. The performance constraint appears to be our lack of knowledge of the higher order auditory processing which underscores a need for more psychoacoustic experiments and electrophysiological investigation of the auditory system. Results of these investigations should generate new speech perception models which perhaps can be applied to solve some of the fundamental problems such as invariance and, ultimately, make computer speech recognition a reality.

References

Alinat, P., "Phonemic Recognition Using a Cochlear Model," in *Spoken Language Generation and Understanding: Proceedings of the NATO Advanced Study Institution* held in Bonas, France, June 26 - July 7, 1979, J. C. Simon (ed.), Dordrecht, Holland: D. Reidel, 1979.

Allen, J., "Magnitude and Phase-Frequency Response to Single Tones in the Auditory Nerve," *Journal of the Acoustic Society of America, 73*, 1983, pp. 2071-2092.

Bladon, R. A. W. and Lindblom, B., "Modelling the Judgement of Vowel Quality Differences," *Journal of the Acoustic Society of America, 69*, 1981, pp. 1414-1422.

Bladon, R.A.W. "Acoustic Phonetics, Auditory Phonetics, Speaker Sex and Speech Recognition: A Thread," in Frank Fallside and William Woods (eds.), *Computer Speech Processing.* London: Prentice-Hall International Ltd., 1985, pp. 29-38.

Blomberg, M., Carlson, R., Elenius, K. and B. Granström. "Auditory Models as Front Ends in Speech Recognition Systems," in *Perkell and Klatt* (eds.), 1986, pp. 108-122.

Blomberg, M., Carlson, R., Elenius, K. and B. Granström. "Experiments with Auditory Models in Speech Recognition," in *Carlson and Granström* (eds.), 1982, pp. 197-201.

Borden, G. J., and Harris, K. S., *Speech Science Primer: Physiology, Acoustics and Perception of Speech*, Baltimore: Williams and Wilkins, 1983.

Carlson R. and Granström, B. (eds.) *The Representation of Speech in the Peripheral Auditory System* Elsevier Biomedical Press, Netherlands, 1982.

Carlson, R. and Granström, B. "Towards an Auditory Spectrograph" in *Carlson and Granström* (eds.), 1982, pp. 109-114.

Chistovich, J. A, Grostrem, M. P., Kozhevnikov, V. A., Lesogor, I. W., Schupljakov, V. S., Taljasin, P. A. and Tjulkov, W. A. "A Functional Model of Signal Processing in the Peripheral Auditory System," *Acustica 31*, 1974, pp. 349-353.

Chistovich, L. A., Lublinskaya, V. V., Malinnikova, T. G., Ogorodnikova, E. A., Stoljarova, E. I., and Zhukov, S. JA. "Temporal processing of peripheral auditory patterns of speech," in *Carlson and Granström* (eds.), 1982, pp. 165-180.

Dolmazon, J-M., Bastet, L. and Schupljakov, V. S., "A functional Model of the Peripheral Auditory System in Speech Processing," *IEEE Acoustic Speech and Signal Processing Rec.*, April, 1977, pp. 261-264.

Dolmazon, J-M., "Representation of Speech-like Sounds in the Peripheral Auditory System in Light of a Model" in *Carlson and Granström* (eds.), 1982, pp. 151-164.

Delgutte, B. "Representation of Speech-like Sounds in the Discharge Patterns of Auditory-nerve Fibers," *Journal of the Acoustic Society of America, 68*, 1980, pp. 843-857.

Delgutte, B. "Some correlates of phonetic distinctions at the level of the auditory nerve," in: *Carlson and Granström* (eds.), 1982, pp. 131-149.

Delgutte, B. "Analysis of French stop consonants using a model of the peripheral auditory system," in *Perkell and Klatt* (eds.), 1986, pp. 163-177.

Delgutte, B. and Kiang, N. Y. S. "Speech Coding in the Auditory Nerve: I. Processing Schemes for Vowel-like Sounds," *Journal of the Acoustic Society of America, 75*, 1984, pp. 866-878.

Durrant, J.D., and Lovrinic, J. H.,*Bases of Hearing Science* Baltimore: Williams and Wilkins, 1984.

Evans, E. F., and Wilson, J. P., "The Frequency Selectivity of the Cochlea," in *Basic Mechanisms in Hearing*, Møller, A. R. (ed.), New York and London, Academic Press, 1973.

Fletcher, H., *Speech and Hearing in Communication*, D. Van Nostrand Co. Inc., Princeton, N. J., 1953.

Flock, Å., "Structure and Function of the Hearing Organ: Recent Investigations of Micromechanics and its Control," in *Carlson and Granström* (eds.), 1982, pp. 43-60.

Godfrey, D. A., Kiang, N. Y. S., and Norris, B. E., "Single Unit Activity in the Posterovenral Cochlear Nucleus of the Cat," *Journal of Comp. Neurology 162*, 1975, pp. 247-268.

Goldhor, R. "A Signal Processing System Based on a Peripheral Auditory Model." Paper No. 28.11 presented at *International Conference on Acoustic Speech and Signal Processing, 1983*, Boston, Massachusettts.

Haggard, M. P. "Simulation and Specification of Peripheral Hearing Impairment" in *Carlson and Granström* (eds.), 1982, pp. 259-264.

Hall, J. L., "Cochlear Models: Two-tone Suppression and the Second Filter," **Journal of the Acoustic Society of America, 67**, 1980, pp. 1722-1728.

Helmoltz, H. L. F., **Die Lehre von den Tonempfindungen als Physiologische Grundlage fur die Theorie der Musik**, first edition, Brunswick: Vieweg, 1863. Eng. trans by A. J. Ellis, 1885.

Houtgast, T. "Auditory-filter Characteristics Derived from Direct-masking Data and Pulsation-threshold Data with a Ripple-noise Masker," *Journal of the Acoustic Society of America 62*, 1977, pp. 409-415.

Hunt, M. J. and Lefebvre, C. "Speech Recognition using a Cochlear Model," *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Tokyo, 1986, pp. 1979-1982.

Itakura, F. "Minimum Prediction Residual Principle Applied to Speech Recognition," *IEEE Transactions of Acoustic Speech and Signal Processing, ASSP-23*, pp.67-72, 1975.

Johnson, D. H., "The Relationship between Spike Rate and Synchronoy in Responses of Auditory-nerve Fibers to Single Tones," *Journal of the Acoustic Society of America, 68*, 1980, pp. 1115-1122.

Johnstone, B. M., and Boyle, J. J. F., "Basilar Membrane Vibrations Examined with the Mossbauer Technique," *Science 158*, 1967, pp. 390-391.

Khanna, S. M. and Leonard, D. G. B., "Basilar Membrane Tuning in the Cat Cochlea," *Science 215*, 1982, pp. 305-306.

Klatt, D. H. "Speech processing strategies based on auditory models," in *Carlson and Granström* (eds.), 1982, pp. 181-196.

Klatt, D. H., "The Problem of Variability in Speech Recognition and in Models of Speech Perception," in *Perkell and Klatt*, 1986, pp. 300-324.

Krasner, M. A. "Critical Band Coder – Digital Encoding of Speech Signals Based on the Perceptual Requirements of the Auditory System," *International Conference on Acoustic Speech and Signal Processing, 1980*, Denver, Colorado, April 1980, MIT Lincoln Laboratories, pp. 327-332.

Leijon, A., "Auditory Models in Hearing Aid Fitting." In *Carlson and Granström* (eds.), 1982, pp. 285-290.

Lindsay, P. and Norman, D., *Human Information Processing: An Introduction to Pscyhology*, Academic Press, Orlando, Florida and London, 1977.

Lyon, R., "A Computational Model of Filtering, Detection, and Compression in the Cochlea," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Paris, May 1982.

Lyon, R., "A Computational Model of Binaural Localization and Separation" Paper No. 24.9, *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Boston, MA, April 14-16, 1983.

Lyon, R., "Computational Models of Neural Auditory Processing," *Proc. IEEE International Conference on Acoustics. Speech and Signal Processing.* San Diego, March, 1984.

Lyon, R. and Lauritzen, N., "Processing Speech with the Multi-Serial Signal Processor," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1985.

Lyon, R. and Lounette, D., "Experiments with a Computational Model of the Cochlea," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Tokyo, 1986.

Martin, F. *Introduction to Audiology*, second edition, Prentice-Hall, Englewood Cliffs, New Jersey, 1981.

Møller, A. R., "Neurophysciological basis for perception of complex sounds," in *Carlson and Granström* (eds.), pp. 43-60.

Moore, B. C. J. *Introduction to the Psychology of Hearing*, second edition, Academic Press, 1982.

Patterson, R. D., "Auditory Filter Shapes Derived with Noie Stimuli," *Journal of the Acoustic Society of America, 59*, 1976, pp. 640-654.

Patterson, R. D. and Nimmo-Smith, I., "Off Frequency Listrengin and Audiotory-filter Assymetry," *Journal of the Acoustic Society of America, 67*, 1980, pp. 229-245.

Patterson, R. D. and Moore, B. C. J. "Auditory Filters and Excitation Patterns as a Representation of Frequency Resolution," in *Frequency Selectivity in Hearing* Academic Press Inc., London, 1986, pp. 123-177.

Perkell, J. S. and Klatt, D. H. *Invariance and Variability in Speech Processes*, Laurence Erlbaum Association, Inc. Hillsdale, New Jersey, 1986.

Pick, G. F., "Level Dependence of Psychophysical Frequency Resolution and Auditory Filter Shape," *Journal of the Acoustic Society of America, 68*, 1980, pp. 1085-1095.

Pickles, J. O *An Introduction to the Pysiology of Hearing*, Academic Press, London and New York, 1982.

Rhode, W. S., "Observations of the Vibration of the Basilar Membrane in Squirrel Monkeys using the Mossbauer Technique," *Journal of the Acoustic Society of America, 34*, 1971, pp. 1218-1231.

Sachs, M. B., Kiang, N.Y.S., "Two-Tone Inhibition in Auditory-Nerve Fibers," *Journal of the Acoustic Society of America, 43*, 1968, pp. 1120-1128.

Sachs, M.B., Young, E.D., "Effects of Nonlinearitites on Speech Encoding in the Auditory Nerve," *Journal of the Acoustic Society of America, 68*, 1980, pp. 858-875.

Sachs, M. B., Young, E.D. and Miller, M.I. "Encoding of Speech Features in the Auditory Nerve," in *Carlson and Granström* (eds.), 1982, pp. 115-130.

Schroeder, M. R., Atal, B. S. and Hall, J. L., "Objective Measure of Certain Speech Signal Degredations Based on the Masking Properties of Human Auditory Perception," in *Frontiers of Speech Communication Research*, Lindblom, B. and Ohman, S. (eds.), London: Academic Press, 1979.

Seneff, S. "Pitch and Spectral Analysis of Speech Based on an Auditory Synchrony Model." (MIT Dissertation), Technical Report 504, MIT Research Lab. of Electronics, Jan. 1985.

Seneff, S. "A Synchrony Model for Auditory Processing of Speech," in *Perkell and Klatt* (eds.), 1986, pp. 115-122.

Shaw, E. A. G. "The External Ear," in *Handbook of Sensory Physiology*, Vol. 5/1, W. D. Keidel and W. D. Neff (eds.), Springer, Berlin, 1974, pp. 445-490.

Small, A.M. "Psychoacoustics" *in Minifie*, F.D., Hixon, T.J. and F. Williams (eds.), *Normal Aspects of Speech, Hearing, and Language.* Englewood Cliffs, N.J.: Prentice-Hall, 1975, pp. 343-420.

Smith, J. C. and Zwislocki, J. J., "Short-Term Adaptation and Incremental Responses of Single Auditory-Nerve Fibers," *Biological Cybernetics 17*, 1975, pp. 169-182.

Syrdal, A.K.; Gopal, H.S. "A perceptual model of vowel recognition based on the auditory representation of American English vowels," *Journal of the Acoustic Society of America, 79* (4), 1986, pp. 1086-1100.

Vogten, L. L. M., "Poure Tone Masking; a New Result from a New Method," in *Facts and Models in Hearing*, E. Zwicker and E. Terhardt (eds.), Springer-Verlag, Berlin, 1974.

von Bekesy, G., *Experiments in Hearing* (edited and translated by E. G. Wever) McGraw-Hill, New York, 1960.

Whitfield, I. C., "Central Nervous Processing in Relation to Spatio-temporal Discrimintation of Auditory Patterns," in *Frequency Analysis and Periodicity Detection in Hearing*, Plomp, R. and Smoorenburg, C. F. (eds.), Sijthoff, Leiden, the Netherlands, 1970.

Zwicker, E. "Subdivision of Audible Frequency Range into Critical Bands (Frequenzgruppen)," in *Journal of the Acoustic Society of America, 33*, 1961, pp. 248-249.

Zwicker, E. and Feldtkeller, R, *Das Ohr als Nachrichtenempfänger*, Stuttgart: S. Hirtzel Verlag, 1967.

Zwicker, E., Terhardt, E. and Paulus E., "Automatic Speech Recognition using Psychoacoustic

Models," *Journal of the Acoustic Society of America, 65*, 1979, pp. 487-498.